ABSTRACT

This invention pioneers an enhanced crawling mechanism and technique called "Enhanced Browser Based Web Crawling". It permits the fault-tolerant gathering of dynamic data documents on the World Wide Web (WWW). The Enhanced Browser Based Web Crawler technology of this invention is implemented by incorporating the intricate functionality of a web browser into the crawler engine so that documents are properly analyzed. Essentially, the Enhanced Browser Based Crawler acts similarly to a web browser after retrieving the initially requested document. It then loads additional or included documents as needed or required (e.g. inline-frames, frames, images, applets, audio, video, or equivalents.). The Crawler then executes client side script or code and produces the final HTML markup. This final HTML markup is ordinarily used for the rendering for user presentation process. However, unlike a web browser this invention does not render the composed document for viewing purposes. Rather it analyzes or summarizes it, thereby extracting valuable metadata and other important information contained within the document. Also, this invention introduces the integration of optical character recognition (OCR) techniques into the crawler architecture. The reason for this is to enable the web crawler summarization process to properly summarize image content (e.g. GIF, JPEG or an equivalent) without errors.